DOCUMENT RESUME

ED 220 478 TM 820 481

AUTHOR Jaeger, Richard M.
TITLE High School Competency Test Standards and the Definition of Competence.
PUB DATE Mar 82
NOTE 16p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New York, NY, March 20-22, 1982).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Competence; *Evaluation Methods; Graduation Requirements; *High Schools; Minimum Competencies; *Minimum Competency Testing; *Standards; Student Certification; Test Validity
IDENTIFIERS Angoff Methods; Jaeger Method; Nedelsky Method; *Standard Setting

ABSTRACT

The implicit definition of competence and the inferential chain that links the standard-setting process to the decision outcomes of the method are considered for two classes of standard-setting procedures: those involving data-free judgments of items and those involving data-based judgment of items. The major underlying assumptions of competence include its existence as a unique construct or trait of adults, that one way to attain competence is through 12 years of formal schooling and graduation, that competence is possessed in greater and lesser degrees, and that it is measurable. The construct definition as a standard for measurement of competence and the validation of standard-setting procedures are discussed in relation to the standard-setting methods of Angoff, Nedelsky and Jaeger. The lack of validation of standard-setting procedures and lack of evidence of construct validity in tests of competence standards are described as factors inhibiting the selection of a method based on superior validity or procedures. (CM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED220478

HIGH SCHOOL COMPETENCY TEST STANDARDS AND THE DEFINITION OF COMPETENCE

Richard M. Jaeger
University of North Carolina - Greensboro

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Richard M. Jaeger

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Presented as part of a symposium entitled "Empirical Evidence and 'Expert' Judgment in Standard Setting: Underlying Assumptions" at the 1982 Annual Meeting of the National Council on Measurement in Education, New York City, March, 1982.

TM 820 481

INTRODUCTION

Although the literature on setting standards for high school competency tests abounds with methods (Hambleton and Eignor cataloged no less than 18 different approaches in their 1980 paper), and several investigators have conducted studies on the comparability of methods (Andrew and Hecht, 1976; Brennan and Lockwood, 1980; Koffler, 1980; Poggio, Glasnapp and Eros, 1981), there has been no systematic analysis of the assumptions underlying various procedures. This paper examines such questions as the implicit definition of competence and the inferential chain that links the standard-setting process to the decision outcomes of the method, for two classes of standard setting procedures: those that involve data-free judgments of items (Angoff, 1971; Nedelsky, 1954; Ebel, 1972) and those that involve data-based judgments of items (Jaeger, 1978; in press).

THE DEFINITION OF COMPETENCE

High school competency testing programs are a product of state legislatures and state boards of education (Gorth and Perkins, 1979a; 1979b). The policy-makers that impose such programs have decided that some students should not receive high school diplomas, even though they have fulfilled all requirements other than successful completion of a competency test. For example, consider the legislation that requires high school competency testing in North Carolina:

"The State Board of Education shall adopt tests or other measurement devices which may be used to assure that graduates of the public high schools and graduates of the nonpublic high schools supervised by the State Board of Education ... possess those skills and that knowledge necessary to function independently and successfully in assuming the responsibilities of citizenship. This article has three purposes: (i) to assure that all high school graduates possess those minimum skills and knowledge thought necessary to function as a member of society, (ii) to provide a means of identifying the strengths and weaknesses in the education process, and (iii) to establish additional means for making the education system accountable to the public for results [General Assembly of the State of North Carolina, Article 39A, Section 115-320.6 'Purpose,' 1977]."

The initiation of a high school competency testing program incorporates a set of assumptions that are sometimes explicit in the enabling legislation and in other cases must be inferred. The major assumptions underlying all legislation that requires the attainment of some minimum score on a competency test in order to receive a high school diploma appear to be as follows:

1) There exists a unique construct, trait, or characteristic of adults, termed competence.

2) Competence is necessary to successful functioning as an adult.

3) One way to attain competence is through 12 years of formal schooling, culminating in high school graduation.

4) Competence is possessed by individuals to a greater or lesser degree; it is not an all-or-nothing variable. Were this not the case, one could not speak of "minimal competence."

5) Society expects holders of a high school diploma to have enough competence to be successful.

6) Competence is measurable -- it is possible to determine whether a given individual has enough of it to function effectively as an adult.

7) A person who does not have enough competence ought not to hold a high school diploma.

COMPETENCE AS A CONSTRUCT

In its fourth meaning, the only one that is relevant to the present discussion, the Oxford English Dictionary defines competence as "Sufficiency of qualification; capacity to deal adequately with a subject." (p. 719). The difficulty of translating this common-language definition into a construct definition has been discussed by Messick (1981). He notes that competence is hardly distinguishable from intelligence, ability, achievement, or proficiency, although these terms embody diverse connotative meanings.

We have noted the necessity of interpreting competence as a construct present in varying amounts if we are to concern ourselves with its compound: minimum competence. Yet the OED definition suggests that competence is a standard, rather than a variable.

Messick notes that competence might be distinguished

from achievement and intelligence by virtue of its connotation of "present capacity." Intelligence implies a prediction of future performance, whereas achievement suggests an accumulation of past performance. We might suggest, therefore, that the construct "minimum competence," in the sense intended by state legislatures, be defined as the smallest amount of presently-held ability necessary to avoid dependence on the state for the necessities of life.

The task of establishing the construct validity of high school competency tests and the methods of setting standards on those tests is not likely to be eased by this definition. However, the requirements of such validations might be more clearly understood.

It is obvious that standards cannot be validated apart from evidence that the competency tests themselves are measures of the construct implied by their enabling legislation. The nomological network (Cronbach and Meehl, 1955; Feigl, 1956) that might be constructed to support the interpretation of a specific score on a competency test, as an indicator of sufficient present ability to avoid dependence on the state for basic necessities of life, could not be established apart from a demonstration that the competency test itself is a measure of ability to sustain societal independence. However, in the discussion that follows, we shall assume that the intended construct interpretations of the test scores themselves have been acceptably validated. To do otherwise would divert this examination of standard-setting methods beyond reasonable

hope of conclusion.

VALIDATION OF STANDARD-SETTING PROCEDURES

Messick (1981) has extended traditional conceptions of test validation to include the required evidential and ethical bases of claims that tests are appropriate to specific uses. He states:

"Whenever test scores are used for a particular purpose or are proposed for such use, they are placed either explicitly or implicitly in a theoretical context of implied relevance and utility. This theoretical context includes not only the nomological network of construct meaning underlying the interpretation of the test but a similar network underlying the meaning of the criterion. Relevance is judged not just in terms of commonality of test content with criterion-task content but in terms of the likely contribution of the test construct to criterion-task performance." (p.10).

The implications of this position for the validation of high school competency tests as measures of competence are profound. Claims to validity can be sustained only by developing a network of construct relationships that includes the criterion as well as the construct underlying the test. Operational measures of all constructs in the networks surrounding the test and the criterion must be developed, and their relationships must survive empirical scrutiny in the face of competing interpretations.

The validation of a method for setting standards is

even more demanding. It is necessary to demonstrate the credibility of claims to the correspondence between status on the criterion construct and status on the test. This requires all of the validity evidence needed to support the use of the test as a measure of competence, in addition to evidence supporting the validity of a test standard as an indicator of minimal competence.

The validation requirements of three methods that have been used for setting standards on high school competency tests will now be discussed. We will consider the operational procedures that define the methods, and the assumptions that link the outcomes of these operations with their implied standard of minimum competence.

Angoff's Method

The standard-setting procedure attributed to Angoff (1971) requires a each of a sample of judges to examine each item on a test to estimate "the probability that a minimally competent examinee would answer the item correctly." The sum of the probabilities estimated by a given judge is defined as the standard recommended by that judge. The average of the recommended standards is the test score adopted as a standard.

Let us examine the assumptions of this process in terms of our earlier construct definitions. First, the procedure demands that each judge conceptualize an individual who possesses the smallest amount of present ability necessary to function independently in sustaining the necessities of life. Having done that, the judge must imagine the ability

of such a person to correctly answer each of a number of test items that typically demand reading or arithmetic skills. An alternative also suggested by Angoff would be to imagine a number of individuals who possess minimal capablities for independent functioning, and to estimate, for that group, the proportion who could correctly answer each of the test items.

The competence criterion is quite general, since the inference applies to contexts that vary in terms of situation and opportunity, and to characteristics of the minimally competent person other than ability. Motivation, initiative, social status, persistence, and discipline are just a few of the constructs that would undoubtedly appear in the nomologic network that framed the competence construct. All of these, and others, would have to be considered by a thoughtful judge in characterizing a minimally competent individual. Perhaps the judge must form mental estimates of success probabilities on items for minimally competent individuals at each of a variety of levels on these constructs, and in a variety of contexts, and average these probabilities in developing each response.

One could argue that the operations of the standard-setting method are not subject to validation, since the resulting inference is based solely upon the final result. In that case, attention could be restricted to examinations of the correspondence between resulting test standards and claims to societal competence. Procedural

validation would require evidence of the credibility of truth claims across a wide variety of standard-setting contexts and in the face of a multiplicity of competing criterion interpretations. Demonstrating that the method resulted in a classification of examinees that accurately distinguished between those who were societally independent and those who were dependent would not suffice. The validity claim necessarily goes beyond any specific sample of examinees, any specific competency test, and any specific sample of judges. For the method to be judged valid, its results must hold in the face of experimental scrutiny that systematically varies each element of its method within the bounds of their definition, and purposefully alters the contexts in which the method is applied. For example, Angoff's procedure does not prescribe the methods by which judges should be selected, the minimum number of judges that should be used, nor the qualifications of judges. These factors would have to be varied systematically to support a claim to validity. The method would have to be applied to samples of examinees that were selected purposefully to represent ranges of performance on variables related to the competency test. As is always true in construct validation, the process is never-ending and asymmetric. No amount of evidence provides proof. Increasing confidence and understanding of generalizability is gained through the accumulation of mutually supportive results. Only one counter-example is needed to repudiate a truth-claim.

Nedelsky's Method

A second popular method of establishing standards on high school competency tests was developed by Nedelsky (1954). In this procedure, each of a sample of judges estimates for each multiple-choice item on a test, the number of options that a minimally competent individual should be able to eliminate as obviously incorrect. The standard recommended by a judge is computed as the sum of the reciprocals of the numbers of options remaining for each item, after obviously incorrect choices have been eliminated. The final standard is taken to be the average of the standards recommended by individual judges.

At the level of analysis of its operations, validation of the Nedelsky method requires an examination of many of the inferences associated with Angoff's method. Once again, judges must conceptualize one or a sample of minimally competent individuals. They must also hypothesize the performance of such individuals on each of a set of test questions likely to require reading or arithmetic skills. However, the inferences demanded by Nedelsky's method are more specific and detailed than those of the Angoff method. Here each judge must consider each test item option and decide whether a minimally competent individual will, or will not, know whether it is incorrect. In validating the method, evidence of its success would have to be amassed across samples of item options, as well as samples of test items.

Otherwise, the arguments advanced in considering the

validity of Angoff's method would appear to apply to Nedelsky's method as well. The nomological network surrounding the criterion would be identical, as would the network surrounding the test. The same counterclaims to generalization and accuracy could be advanced. The experimental procedures involved in their examination would likely be identical, apart from purposeful variation of the relative difficulties of options associated with various item stems.

Jaeger's Method

The third standard-setting method we shall consider briefly is one proposed by Jaeger (1973; in press). In this procedure, each of a sample of judges decides, for each item on a competency test, whether or not every high school graduate in the political jurisdiction imposing the test, should be able to answer the item correctly. The standard recommended by a given judge is taken to be the number of items for which the judge makes a positive decision. The final standard is computed as the average of the standards recommended by all judges. The method requires repeated judgments in the face of a variety of evidence on the past performances of examinees and the consequences of a variety of standards. However, the final standard is still computed as an average of judgments.

At the level of its operations, this standard-setting method is far less demanding than either the Angoff or Nedelsky procedures. Judges are not required to

conceptualize a minimally competent examinee, nor are they called upon to estimate what the performance of such examinees would be, either on individual items or in knowing the correctness of individual item options. It might appear, therefore, that the construct validation of the Jaeger method would be simpler and less demanding. This potential advantage is illusory.

The truth claims resulting from application of Jaeger's standard-setting method are identical to those of Angoff's and Nedelsky's procedures. Construct validation of Jaeger's method would require the same sequence of careful definition, establishment of nomological networks surrounding the test and the criterion, and empirical research as would Angoff's method. Since Jaeger's method specifies a sampling procedure for selecting various types of judges, experiments would involve variation of the types and numbers of judges used to derive test standards, rather than sampling procedures, as would be the case with Angoff's method. Otherwise, the validity of test standards would have to be established across contexts, tests, and characteristics of examinees, just as was true for Angoff's method.

SUMMARY

It is not an exaggeration to say that we presently have no idea of the validity of any procedure used to set standards on high school competency tests. Since validation of our standard-setting procedures requires validation of our competency tests, this is not surprising. We also lack evidence of the construct validity of our competency tests.

It does not appear that a standard-setting procedure can be selected on the basis of its superior claim to validity nor because of the procedural ease with which it could be validated. Details of method appear to make little difference in the requirements for adequate construct validation.

In the face of these conclusions we must either ignore the Standards for Educational and Psychological Tests and continue selecting our procedures in accordance with our satisfaction or the political viability of their results, or suspend the withholding of high school diplomas on the basis of competency test performance until such time as adequate validity evidence is amassed. Our best hope may be that the courts will make that decision for us.

REFERENCES

American Psychological Association, American Educational Reserch Association, & National Council on Measurement in Education. Standards for educational and psychological tests. Washington, DC: American Psychological Association, 1974.

Andrew, B. J. & Hecht, J. T. A preliminary investigation of two procedures for setting examination standards. Educational and Psychological Measurement, 1976, 36, 35-50.

Angoff, W. H. Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational measurement. Washington, D.C.: American Council on Education, 1971.

Brennan, R. L. & Lockwood, R. E. A comparison of the Nedelsky and Angoff cutting score procedures using generalizability theory. Applied Psychological Measurement, 1980, 4, 219-240.

Cronbach, L. J. & Meehl, P. E. Construct validity in psychological tests. Psychological Bulletin, 1955, 52, 281-302.

Ebel, R. L. Essentials of educational measurement. Englewood Cliffs, NJ: Prentice-Hall, 1972.

Feigl, H. Some major issues and developments in the philosophy of science of logical empiricism. In Feigl, H. & Scriven, M. (Eds.), Minnesota studies in philosophy of science: The foundations of science and the concepts of psychology and psychoanalysis. Minneapolis: University of Minnesota Press, 1956.

General Assembly of the State of North Carolina. Article 39A, Section 115-320.6, "Purpose," 1977.

Gorth, W. P. & Perkins, M. R. A study of minimum competency testing programs. Final program development resource document. Amherst, MA: National Evaluation Systems, Inc., December, 1979a.

Gorth, W.P. & Perkins, M. R. A study of minimum competency testing programs. Final typology report. Amherst,MA: National Evaluation Systems, Inc., 1979b.

Jaeger, R. M. An interative, structured judgment process for setting standards on competency tests: theory and application. Educational Evaluation and Policy Analysis. (in press).

Jaeger, R. M. A proposal for setting a standard on the North Carolina High School Competency Tests. Presented at the Spring Meeting of the North Carolina Association for

Research in Education, Chapel Hill, 1978.

Koffler, S. L. A comparison of approaches for setting proficiency standards. Journal of Educational Measurement, 1980, 17, 167-178.

Messick, S. Evidence and ethics in the Evaluation of Tests. Research Report RR-81-9, Princeton, NJ: Educational Testing Service, May, 1981.

Nedelsky, L. Absolute grading standards for objective tests. Educational and Psychological Measurement, 1954, 14, 3-19.

Poggio, J. P., Glasnapp, D. R. & Eros, D. S. An empirical investigation of the Angoff, Ebel, and Nedelsky standard-setting methods. Presented at the Annual Meeting of the American Educational Research Association, Los Angeles, April, 1981.

The Oxford English Dictionary. Oxford, England: Oxford University Press, 1971.